**Sujayeendra**

**Senior Data Engineer Phone:** (413) 238-1517

 **Email: sujayeendrak@gmail.com**

**PROFESSIONAL SUMMARY:**

* **8+ Years** of experience in Data Engineer, Data modeling, and ETL Development proficient in Python, Apache Hadoop Ecosystem like HDFS, MapReduce, Hive, Sqoop, Oozie, HBase, Spark-Scala, Kafka and Big Data Analytics, AWS, GCP and Azure.
* Excellent understanding experience of Hadoop architecture and various components such as **HDFS, Job Tracker, Task Tracker, Name Node, Data Node,** and **MapReduce** programming paradigm.
* Expertise in **Big Data technologies, Data Pipelines, SQL,** and **Cloud-based RDS,** Distributed Database, Serverless Architecture, Data Mining, Web scraping, Cloud technologies like AWS EMR, and Cloud Watch.
* Hands-on experience in ingesting data into Data Warehouse using various data-loading techniques.
* Strong Knowledge in **Data Validation, Data Cleansing, Data Verification, Data Profiling, Integration**, and Master Data Management Services.
* Working experience on Hadoop Cluster architecture and sources and involved in HDFS maintenance and loading of structured and unstructured data.
* Experience in managing and reviewing **Hadoop log files** and NoSQL databases like **HBase**, and **MongoDB.**
* Experienced in using distributed computing architectures and building pipelines using AWS products such as **(EC2, Redshift, EMR, Elastic search, EBS, S3), Hadoop, Python, Spark, and effective use of MapReduce, SQL,** and **Cassandr**a to solve big data type problems.
* Working capable of using AWS utilities such as EMR, S3, and cloud watch to run and monitor Hadoop and spark jobs on Amazon Web Services (AWS).
* Strong experience in creating and monitoring Hadoop clusters on **AWS EC2, VM, Hortonworks Data Platform, CDH3, and CDH4 Cloudera Manager on Linux, and Ubuntu OS.**
* Experience with **Azure Cloud Services, SQL Azure, Azure Analysis, Azure Monitoring, Azure data factory.**
* Experience in GCP with compute engine, cloud Dataflow, Cloud Dataproc, Big Query, cloud load balancing, cloud storage, cloud SQL, GKE clusters, cloud deployment manager, GSUTIL, and Pub/Sub cloud shell.
* Well-versed in Data Migration, Data Conversions, and Data Extraction/ Transformation/Loading (ETL).
* Good experience using Apache NiFi to automate data movement between various Hadoop systems.
* Experience in importing and exporting data using Sqoop from HDFS to Relational Database Systems.
* Experience in writing complex SQL, PL/SQL, and T-SQL queries implementing procedures, and functions, and improving the performance of databases.
* Good understanding of Web services, SOAP, REST API, and WSDL.
* Experience in cluster automation using Shell Scripting and designing data platforms using AWS, GCP for Data Analysis, and Machine learning.
* Good experience working on analysis tools like **OBIEE, and Tableau**.
* Worked to design complex data models both in real-time and offline analytic processing and provided support for data profiling and data quality functions.
* Worked with Datadog for anomaly detection and configure and tuned the features to ensure the right alerts are triggered when there is an issue.
* Executed medium-to-complex SQL queries for data analysis and data validation.
* Experience in developing **CI/CD** (**continuous integration** and **continuous deployment**) and automation using **Jenkins, Git, docker, and Kubernetes** for ML model deployment.
* Worked on several Python modules such as **Numpy, Pandas, Matplotlib, Scikit-learn, and Pyspark** module

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **BigData /Hadoop** | MapReduce, Spark, Spark SQL, Azure, Spark Streaming, Kafka, PySpark, Pig, Hive, Sqoop, HBase, Flume, Yarn, Oozie, Zookeeper, HDFS, Cloudera. |
| **Languages** | Python (NumPy, SciPy, Pandas, Genism, Keras), Java Script, Shell Scripting, SQL. |
| **NOSQL DB** | Cassandra, HBase, MongoDB, MariaDB. |
| **Dev Tools** | Microsoft SQL Studio, IntelliJ, Azure Databricks, Eclipse, Net Beans. |
| **Cloud** | EC2, IAM, S3, Autoscaling, CloudWatch, Route53, EMR, RedShift, RDS, GCP, Lambda,  |
| **Build Tools** | Jenkins, Toad, SQL Loader, PostgreSQL, Talend, Maven, ANT, RTC, RSA, Oozie, Hue, SOAP UI |
| **Reporting Tools** | MS Office (Word/Excel/Power Point/ Visio/Outlook), Crystal reports XI, SSRS, Cognos,Datadog. |
| **Databases** | MS SQL Server, MySQL, Oracle, DB2, Teradata, Netezza,Spark SQL, Snowflake, Mango DB. |
| **Operating Sys** | Windows, UNIX, LINUX, Sun Solaris, PowerShell. |

**PROFESSIONAL EXPERIENCE**

**Client: Sam's Club, GA Mar2021 - Present**

**Sr Data Engineer**

**Responsibilities:**

* Create data capture process, organize data, and normalize data for analysis, insights, and real-time passenger information.
* Hands-on Experience with Amazon EC2, Amazon S3, Amazon Redshirt, Amazon Elastic Map Reduce, Auto Scaling, Load Balancing
* Integrate data from many applications and systems using automation platforms for DBMS.
* Developed an automation tool using Python libraries like Pandas, Matplotlib, NumPy
* Performed Data profiling, Validation and Integration, Data Migration, and Data distribution testing.
* Proficient in Preparing Test Plans, Test Scripts, Test Scenarios, Test Designs, and Test Cases for both Manual and automation testing.
* Skilled in creating and maintaining test data sets for ETL testing, including synthetic data generation, data masking, and data subsetting techniques, while ensuring data privacy and compliance with data protection regulations.
* Strong knowledge of PySpark's SQL and DataFrame API, enabling seamless integration with structured and semi-structured data sources, performing complex data manipulations, and executing advanced analytics and machine learning algorithms.
* Analyze obtained data and validate them using different operational concepts and evaluate any problems if present greatly involving Python and IOT Devices
* Extensively used AWS S3, EC2, and EMR instances to deploy and test the applications in various environments (DEV, QA, PROD).
* Implemented and managed event-driven architectures using AWS EventBridge to enable real-time data processing and event-driven workflows.
* Designed and developed EventBridge rules and event patterns to route and transform data between various AWS services, such as Lambda functions, SNS topics, SQS queues, and Kinesis streams.
* Developed Python scripts to create EMR, add steps, and delete stacks.
* Proficient in designing and executing complex ETL test cases, scenarios, and scripts to validate data extraction, transformation, and loading processes, adhering to industry best practices and regulatory requirements.
* Built Amazon ECS Cluster, integrated the data from all sources, and created a common schema.
* Developed Sqoop jobs to load data from RDBMS to external systems like HDFS and Hive and vice versa.
* Strong knowledge of pharmacy domain data, including medication data, patient records, healthcare claims, and prescription data, enabling comprehensive and accurate testing of ETL workflows specific to the industry.
* Developed Spark applications using PySpark and SparkSQL for data extraction, transformation, and aggregation from multiple file formats.
* Used HiveQL to perform data analysis on the data stored in Hive according to business requirements.
* Designed and developed Spark code using Spark SQL for high-speed data processing to meet critical business requirements.
* Managed and scheduled several jobs to run overtime on the Hadoop cluster using Oozie

**Environment:** Hadoop, HDFS, Spark, Sqoop, Oozie, Hive, Python, AWS.

**Client: T-Mobile, KS Sep 2019 – Feb 2021**

**Sr Data Engineer**

**Responsibilities:**

* Worked on publishing interactive data visualization dashboards using Tableau and auto deploy content using GIT/Bitbucket
* Analyzed DB discrepancies and synchronized the Staging, Development, UAT, and Production DB environments with data models.
* Implemented Proofs of Concept on Hadoop stack and different big data analytic tools, migration from different databases to Hadoop.
* Read the data from the complex JSON string from the snowflake table and keep them in different tables.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the SQL activity
* Built a Data Cluster on Confidential Azure to extract actionable insights for data collected from IOT sensors installed in excavators.
* Access the data from Redshift and combine the data with Snowflake to define the relationship between data sources in a query.
* Involved in making Hive tables, stacking information, composing Hive inquiries, and producing segments and basins for enhancement.
* Extracted data from SQL Server Database and Teradata copied into HDFS File system and used Hadoop tools such as Hive to retrieve and analyze the data required for building models.
* Perform validation and testing of models to ensure adequacy and accuracy of performance using Excel and MS Access
* Took ownership of Source to Target Mapping and tracked and maintained changes.
* Performed Extraction, Transformation, and Loading (ETL) using Informatica power center.
* Data access through extensive use of Excel (HLOOKUP, VLOOKUP, and Pivot tables) for in-depth analysis
* Worked in a Scrum Agile process & Writing Stories with two-week iterations delivering product for each iteration

**Environment**: MySQL, Hadoop, Python, R and MS Office suite, Windows Power Shell, MS Access, Snowflake

**Client: Sanofi, MA Dec 2017 – Aug 2019**

**Data Engineer**

**Responsibilities:**

* Designed and Developed Enterprise data lake which allows various data types of data from multiple data resources.
* Designed and implemented a real-time data pipeline to work with semi-structured data by incorporating a larger volume of raw records from multiple data sources using **Kinesis Data Steam** and **kinesis firehouse.**
* Created **AWS Lambda function** to read from the Producer and write records to an **Amazon DynamoDB** table as they arrive.
* Worked with **EMR** to transform and move big data into AWS data stores and databases in **S3 and DynamoDB.**
* Created and launched **AWS EC2** instances to execute jobs on **EMR** to store the results in **S3.**
* Designed and set up **Enterprise Data Lake** to provide support in multiple areas such as Analytics, processing, storing, and reporting of big and rapidly changing data.
* proficient in working with a variety of data pipeline technologies, including Apache Spark, Apache Kafka, Apache NiFi, or Talend, to ensure seamless data integration and movement across different data sources and systems.
* Worked with Hadoop distributed file system (**HDFS), S3 Storage**, and big data formats like **parquet, and JSON.**
* Configured AWS Redshift clusters, AWS Redshift spectrums for querying, and **AWS Redshift** Data share for transferring the data among clusters.
* Automated the ETL processes using **Apache Airflow** and **S3** for storing the data in batches.
* Developed Spark Applications by using Python and Implemented **Apache Spark** data processing Project to handle data from various **RDBMS** and Streaming sources.
* Developed Python code to satisfy requirements and perform data processing and analytics using inbuilt libraries.
* Created Lambda functions with Boto3 for unused **AMIs** in all application regions to reduce the cost of EC2 resources.
* Worked with Spark for improving the performance and optimization of the existing algorithms.
* Deployed Applications on **AWS EC2** instances and configured the storage on **S3 buckets.**
* Used **Snowflake** to integrate data from various sources, including on-premises databases, cloud-based applications, and data lakes.
* Configured **S3 buckets** with policies to automatically archive the infrequently accessed data to storage classes.
* Involved in Analyzing raw files from **S3 data lake** using **AWS Athena, and Glue** without loading the data in the database.
* Implemented **IAM** roles for various resources like **EC2, S3, and RDS** to communicate with each other.
* Worked with **Datadog** for anomalies detection and configure and tuned the features to ensure the right alerts are triggered when there is an issue.
* Analyzed complex data and identified anomalies, trends, and risks to provide insights to improve internal controls.

**Environment:** AWS, Hadoop, Spark, AWS Kinesis, Parquet, JSON, SNS, Redshift, Apache Airflow, Snowflake, MYSQL, EC2, S3, AWS Athena, Glue.

**Client: Ceequence Technologies, Hyderabad, India Feb 2015 – Sep 2017**

**Hadoop Developer**

**Responsibilities:**

* Engaged with different trams in gathering requirements to design ETL migration process from Existing RDBMS to Hadoop Cluster using Sqoop.
* Proficient in using testing frameworks and tools such as Apache Hadoop MRUnit, Apache HiveTestUtils, Apache PigUnit, or Apache Spark Testing Base, to automate the testing of Hadoop applications, Hive queries, Pig scripts, and Spark jobs.
* Worked to develop HIVE Queries for Data Transformation and Data analysis.
* Developed, implemented, and tested Python-based web applications interacting with MySQL.
* Developed skills in how to load data from Hive table to MySQL and DB2 using Python scripts.
* Involved in web services and Hibernate in a fast-paced development environment.
* Performed ETL operations using Informatica, PL SQL, and UNIX shell scripts and worked with SSIS.
* Developed strong ETL design and development skills and understood the tradeoffs of various design options on multiple platforms using multiple protocols.
* Developed ETL pipelines in and out of the data warehouse using a combination of Python and Snowflakes.
* Worked with SQL Server Reporting Services (SSRS) and created various types of reports like Parameterized, Cascading, Conditional, Matrix, Table, Chart, and Sub Reports.
* Defined static and dynamic repository variables to modify metadata content to adjust to a changing data environment.
* Researched Lean Six Sigma principle and implemented an online portal that replaced the Excel sheets and optimized the storage of data by reducing.
* Gained in-hands experience in agile software development process including analysis design and worked on Scrum/Jira and Confluence.
* Gained experience with DataStax Spark connector to store the data in the Cassandra database and also to get the data from the Cassandra database.

**Environment**: Hadoop, MapReduce, HDFS Hive, Java Hadoop distribution of Horton Works, Cloudera, Pig, HBase, Linux, XML, MySQL, MySQL Workbench Java 6, Eclipse, Oracle 10g PL/SQL, SQL PLUS Sub Version Cassandra.